Rescaling Clustering Trees Using Impact Ratios for Robust Hierarchical Speaker Clustering

Houman Ghaemmaghami, David Dean, Shahram Kalantari, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia

{houman.ghaemmaghami, d.dean, s1.kalantari, s.sridharan}@qut.edu.au

Abstract

We present a novel method for improving hierarchical speaker clustering in the tasks of speaker diarization and speaker linking. In hierarchical clustering, a tree can be formed that demonstrates various levels of clustering. We propose a ratio that expresses the impact of each cluster on the formation of this tree and use this to rescale cluster scores. This provides score normalisation based on the impact of each cluster. We use a state-of-the-art speaker diarization and linking system across the SAIVT-BNEWS corpus to show that our proposed impact ratio can provide a relative improvement of 16% in diarization error rate (DER).

Index Terms: speaker diarization, speaker clustering, cluster impact ratio

1. Introduction

The rapid increase of multimedia archives has brought about a need for techniques that can be used to automatically annotate and index large sets of recordings with respect to speaker identity. To do this, it is necessary to first carry out speaker diarization to reveal *'Who spoke when?'* in each recording [1, 2, 3]. If the speakers appear across multiple recordings, speaker linking can then be applied to find instances of recurring identities between recordings [4, 5, 6, 7]. Applying speaker diarization and speaker linking to an archive of recordings will thus provide information regarding the number of unique speakers in each recording and when they speak, as well as the unique number of speakers across the entire archive and the recordings that each appear in. In this paper we will use the term *speaker attribution* to refer to the combined tasks of diarization and speaker linking.

Speaker attribution has been an important source of information for multimodal person recognition in multimedia datasets [5, 8, 9]. An increase in volume of such datasets can severely degrade the efficiency of traditional systems, which commonly use computationally expensive hierarchical cluster merging and re-training schemes [2, 3]. It is thus necessary to utilise an efficient (yet robust) speaker modeling and clustering approach that can overcome such problems. Recent advances in speaker recognition have provided robust speaker modeling techniques such as i-vector modeling [10] and joint factor analysis (JFA) modeling with session compensation [11, 12]. These methods have been the most popular techniques in speaker attribution research [6, 7, 13]. In order to efficiently cluster the speaker models without retraining, we have previously proposed using complete-linkage clustering [7, 14], based on the pairwise cross-likelihood ratio (CLR) similarity metric calculated between JFA adapted models [15, 3], and have used this technique to conduct speaker attribution. We will thus use this state-of-the-art speaker attribution system as baseline [16], to evaluate our proposed techniques.

In this paper we propose an approach for achieving robustness in hierarchical (or linkage) speaker clustering. In linkage clustering the initial cluster nodes are chosen based on the highest pairwise cluster similarity scores (or lowest pairwise distances). The scores between these nodes and other clusters are then updated based on a linkage rule [17], without the need for model retraining. This approach provides a clustering tree that maps out every level of the hierarchical clustering process. A stopping criterion can then be applied to cut this tree at a level that would provide the most suitable clustering decision. Before selecting an appropriate level of clustering, we propose using the full clustering tree to obtain (for each cluster) a measure of the impact of that cluster on the formation of the clustering tree. We call this the *cluster impact ratio* and use this ratio to rescale each cluster's set of pairwise scores before making a clustering decision. We hypothesise that this approach can achieve robustness in hierarchical speaker clustering through normalising pairwise cluster scores with respect to their impact on the entire clustering process. We evaluate our clustering approach by applying this technique to our speaker attribution system across the SAIVT-BNEWS corpus of Australian broadcast data [16]. We demonstrate a relative improvement of 16% in diarization error rate (DER) over our baseline performance, as well as improvements to the cluster purity and coverage metrics. In addition, we show that our approach provides a better estimate of the unique number of speakers (compared to the baseline system) across this corpus.

2. Baseline speaker attribution

Throughout this paper we employ our state-of-the-art speaker attribution system [16], as baseline. This system has been shown to be robust across multiple audio domains and efficient for processing large datasets [16, 7, 15]. We provide a brief description of this system. As we are attempting to improve speaker clustering, we begin by presenting the speaker modeling and clustering scheme used in our baseline system. After that we describe the baseline speaker diarization and speaker linking modules.

2.1. Speaker modeling and clustering

The baseline system uses JFA modeling with session compensation [18, 12], which makes it ideal for modeling and comparing a variety of speakers across different session conditions. In this method of speaker representation, a universal background model (UBM) is used to obtain a constrained offset of the speaker- and session-independent Gaussian mixture model (GMM) mean supervector, $\boldsymbol{m}$,

$$\boldsymbol{m}_i(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(s) + \boldsymbol{D}\boldsymbol{z}(s) + \boldsymbol{U}\boldsymbol{x}_i(s), \qquad (1)$$

where s is the speaker index, $\boldsymbol{m}_i(s)$ is the speaker-dependent, session-dependent mean supervector of dimension $CL \times 1$. C is the number of mixtures used in UBM training and L is the dimension of the extracted acoustic features. $\boldsymbol{x}_i(s)$ is a low-dimensional representation of variability in session i, and $\boldsymbol{U}$ is a low-rank transformation matrix from the session subspace to the UBM supervector space. $\boldsymbol{y}(s)$ is the speaker factors that represent a speaker in a specified subspace with a standard normal distribution [12]. $\boldsymbol{V}$ is a low-rank transformation matrix from the speaker subspace to the UBM supervector space, and $\boldsymbol{Dz}(s)$ is the residual variability not captured by the speaker subspace. We train the JFA hyperparameters using the coupled expectation-maximization (EM) algorithm proposed by Vogt et al. [12].

After JFA adapted speaker models are obtained, a pairwise cross-likelihood ratio (CLR) similarity score is computed between all models. Given two speakers/clusters i and j, and their corresponding feature vectors $\boldsymbol{q}_i$ and $\boldsymbol{q}_j$, respectively, their pairwise CLR score α_{ij} is computed as,

$$\alpha_{ij} = \frac{1}{K_i} \log \frac{p(\boldsymbol{q}_i|M_j)}{p(\boldsymbol{q}_i|M_B)} + \frac{1}{K_j} \log \frac{p(\boldsymbol{q}_j|M_i)}{p(\boldsymbol{q}_j|M_B)}, \quad (2)$$

where, K_i and K_j represent the number of observations in $\boldsymbol{q}_i$ and $\boldsymbol{q}_j$, respectively. M_i and M_j are the adapted models, and $p(\boldsymbol{q}|M)$ is the likelihood of $\boldsymbol{q}$, given model M, with M_B representing the UBM. The CLR has been shown to be a robust metric for comparing speaker models [3]. Based on our previous work [15, 16], employing the CLR metric in this manner appears to provide a natural comparison threshold value of 0.0. We will thus use this threshold value.

Speaker clustering is carried out based on the pairwise CLR scores and using complete-linkage clustering. Complete-linkage is a form of hierarchical clustering that employs a linkage rule to update the pairwise cluster scores after a merge takes place [17]. Complete-linkage can be carried out without a retraining stage, using only the initial CLR scores [14]. This makes it highly efficient for processing clustering large sets of data. In this clustering approach, the most similar pairs of clusters (with highest CLR score) are merged to form starting nodes. The pairwise score between new clusters and remaining clusters is then updated to reflect the CLR score between their most dissimilar elements (lowest CLR score). For example, if we merge two clusters C_i and C_j into $C_{i'} = \{C_i, C_j\}$, the score between the newly formed cluster $C_{i'}$ and any remaining cluster C_x will be $\alpha_{i'x}$, where,

$$\alpha_{i'x} = \min(\alpha_{ix}, \alpha_{jx}). \quad (3)$$

This merge and update scheme is repeated until there are no CLR scores that are above the threshold value of 0.0. Complete-linkage clustering thus takes into account the worst-case scenario by pessimistically updating the scores that could link clusters to one another in future merges. This is a desirable characteristic that provides a cautious method of clustering, which can be carried out with efficiency when processing large datasets. In addition, this approach has been shown to outperform traditional clustering with retraining and other state-of-the-art techniques employed for speaker clustering in the task of speaker attribution [15, 7].

2.2. Speaker diarization and linking

The speaker diarization module of the baseline system is used to annotate independent recordings with respect to speaker identity [16]. This system uses the hybrid voice activity detection (VAD) approach proposed for the ICSI RT-07 system [2]. A linear segmentation of the audio is carried out, using an ergodic hidden Markov model (HMM) with Viterbi segmentation [16], to obtain a set of speaker change points. These segments are then modeled and clustered using the approach in Section 2.1, followed by HMM/Viterbi resegmentation to refine the obtained speaker utterance boundaries. This clustering and boundary refinement stage is repeated once more to ensure no speakers/clusters are left behind.

The baseline speaker linking module is initialised using the diarization output. This module is responsible for linking intra-recording speakers across independent recordings. This is carried out using the speaker modeling and clustering process detailed in Section 2.1. The accuracy of the linking module is thus highly dependent on the output of the diarization stage.

3. Cluster impact ratio

In hierarchical (or linkage) clustering, distance metrics are commonly employed to represent the pairwise relationship between participating clusters [17]. These pairwise distances can be used to construct a clustering tree, otherwise known as a dendrogram, which demonstrates every level of the hierarchical clustering process. An example of a clustering tree is shown in Figure 1. In this example, the dendrogram is depicting the hierarchical clustering of five clusters to one cluster. Every time a merge takes place between a pair of clusters, the two clusters that have been merged are joined using a coloured upside down U-shaped connection, with cluster indices shown on the x-axis and pairwise distances displayed on the y-axis.

In linkage clustering, pairwise score updates (after a cluster merge) are conducted based on a linkage rule and without the need for model retraining [17]. For this reason, when given N clusters it is possible to efficiently construct a clustering tree (similar to that in Figure 1) that represents all possible clustering levels. In a linkage clustering tree, the formation of the higher levels depends on the outcome of prior levels in the tree. This means that, in the example given in Figure 1, the initial cluster nodes formed at level $N = 1$ are more influential on the final outcome of the clustering process than those at level $N = 4$. We propose a ratio that would express the impact of each participating cluster on the final clustering tree without taking into account the pairwise scores between clusters. We call this ratio the cluster impact ratio (CIR) and calculate the CIR for a cluster C_i, in an N level linkage clustering tree as,

$$\lambda_i = \frac{1 + (N - N_i)}{N}, \quad (4)$$

where λ_i is the CIR for cluster C_i, N is the total number of clustering levels and N_i is the level at which cluster C_i is first merged with another cluster. For example, in Figure 1 the CIR for cluster C_1 is $\lambda_1 = 0.8$.

We propose using the CIR for each cluster to rescale the pairwise cross-likelihood ratio (CLR) scores for that cluster prior to speaker clustering. We hypothesise that this achieves a form of score normalisation based on the impact of each cluster on the complete-linkage clustering process, which would allow for more robust speaker clustering. In order to incorporate our proposed CIR in to the complete-linkage clustering process, we propose rescaling the original pairwise CLR scores before making a final clustering decision. We achieve this by revising (2):

$$\alpha'_{ij} = \frac{\lambda_i}{K_i} \log \frac{p(\boldsymbol{q}_i|M_j)}{p(\boldsymbol{q}_i|M_B)} + \frac{\lambda_j}{K_j} \log \frac{p(\boldsymbol{q}_j|M_i)}{p(\boldsymbol{q}_j|M_B)}, \quad (5)$$

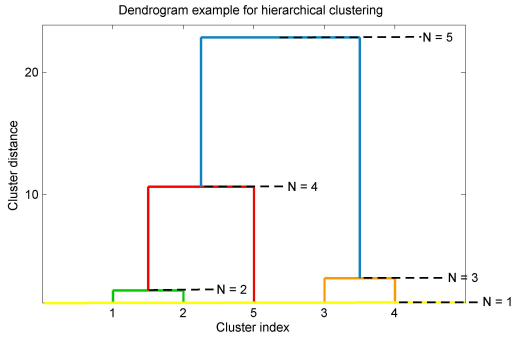

Figure 1: *An example of a dendrogram (clustering tree) with N indicating the level of clustering out of a total of five levels.*

where α'_{ij} is the rescaled pairwise CLR score between clusters i and j, and λ_i and λ_j are the CIR values for clusters i and j, respectively. It can be seen that this rescaling process can now be repeated through achieving a new clustering tree based on the rescaled α'_{ij} scores. This allows for the iterative application of CIR to the complete-linkage clustering process, which we will investigate in Section 4. It is vital to note that our proposed CIR value is only appropriate for rescaling similarity scores. When dealing with distance scores, it is necessary to employ $(1 - \lambda_i)$, in place of λ_i as the CIR for cluster C_i.

4. Evaluations

For evaluation, we employ the SAIVT-BNEWS evaluation corpus [16], which is a publically available collection of Australian broadcast television data. This corpus contains 55 broadcast television program videos, most of which are broadcast news programs with inter-related topics that allow for recurring speaker identities across multiple recordings and session conditions. This dataset provides a large variety of speakers, such as reporters, politicians, presenters, children and elderly people. The 55 files in the SAIVT-BNEWS dataset range from 47 seconds to 5 minutes and 47 seconds, contain from 1 to a maximum of 9 unique speakers within each recording, with a total of 92 globally unique speaker identities across the dataset. The dataset is also provided with a set of reference annotation labels that can be used for calculating the evaluation metrics. This makes the SAIVT-BNEWS corpus a suitable dataset for conducting speaker attribution research.

We use the standard diarization error rate (DER) [19], cluster purity (CP) and cluster coverage (CC) [14], as our evaluation metrics. Ideally, we would like to minimise the DER, while maximising the CP and CC metrics. We first evaluate our baseline speaker attribution system by carrying out speaker diarization and then speaker linking. We then apply our proposed speaker clustering approach with CIR to the baseline system and compare results. We refer to the speaker diarization error, which reflects the overall within-recording errors, as DER. To distinguish between the diarization error rate and the error associated with speaker attribution (diarization and linking), we use the term attribution error rate (AER). AER is in fact the DER measure computed within and between all recordings, thus taking into account recurring identities across multiple recordings.

We employ 20 MFCC features, including the 0^{th} order coefficient, extracted using a 20 bin Mel-filterbank, 32 ms Hamming window and a 10 ms window shift to conduct speaker segmentation and VAD. For speaker modeling, we use 13 MFCC features, including the 0^{th} order coefficient and deltas, extracted in the same manner, with feature warping [20]. We use a combined-gender UBM of 512 mixtures, with a 50-dimensional session and 200-dimensional speaker subspace trained on NIST SRE 2004 and Switchboard II (phase 2 and 3) [21].

Table 1: Baseline performance improves with CIR rescaling.

Diarization	**DER**	**CP**	**CC**	**Speakers**
Baseline	13.2%	80.8%	92.6%	166
Baseline+CIR	13.0%	81.1%	92.7%	166
Attribution	**AER**	**CP**	**CC**	**Speakers**
Baseline	35.9%	74.6%	74.9%	67
Baseline+CIR	**32.3%**	**75.3%**	**76.8%**	**75**

4.1. Experimental results

We evaluated the baseline system before and after CIR rescaling. For consistency with the baseline system, we use the CLR clustering stopping threshold of 0.0 throughout our experiments to conduct complete-linkage speaker clustering. Table 1 displays the results of the baseline system across the SAIVT-BNEWS dataset, before and after applying our proposed CIR rescaling approach to hierarchical speaker clustering. It must be noted that the number of unique speakers obtained using diarization represents the sum of the number of unique intra-recording speaker identities, which will typically be higher than the true 92 globally unique inter-recording speaker identities.

From Table 1, it appears that CIR rescaling has a minimal effect on the diarization outcome. This is apparent from the little improvement observed with respect to the DER, CP and CC metrics. This is while applying CIR rescaling to the baseline speaker linking module results in noticeable improvement of the attribution performance. It can be seen that the attribution accuracy has increased as the AER has been reduced, while both CP and CC metrics have increased. In addition, we now obtain 75 unique speaker identities, which is closer to the true number of 92 unique speakers. This indicates that the CIR rescaling has reordered the linkage clustering tree, achieving a more robust speaker clustering solution than the baseline approach.

CIR rescaling does not impact speaker clustering at the diarization level to the same extent as it does speaker linking. We believe this to be due to the fact that in diarization we are often concerned with clustering short speaker segments that cannot be modeled reliably. In addition, the final Viterbi resegmentation stage of the diarization module can overshadow any effect that CIR rescaling may have on segment clustering at the diarization level. For this reason, we will continue our investigation by conducting diarization using the baseline system and only applying CIR rescaling to the baseline speaker linking module.

As stated in Section 3, CIR rescaling may be applied in an iterative manner. We apply 10 iterations of CIR rescaling to the speaker linking module of the baseline system. Figure 2 displays the performance of the baseline system at each iteration of CIR rescaling, where iteration 0 indicates the baseline system performance without CIR, iteration 1 is the baseline with 1 iteration of CIR applied to the clustering process of the baseline speaker linking module, iteration 2 is the baseline system with 2 iterations of CIR at the linking module and so on. It can be seen that the best performance is achieved at iteration 4, with little change observed beyond the 4^th iteration of CIR rescaling. The best system performance at iteration 4 is presented in

Table 2: Baseline performance improves further through applying iterative CIR rescaling (4 iterations).

Attribution	AER	CP	CC	Speakers
Baseline	35.9%	74.6%	74.9%	67
Baseline+4(CIR)	**30.1%**	**76.7%**	**78.3%**	**82**

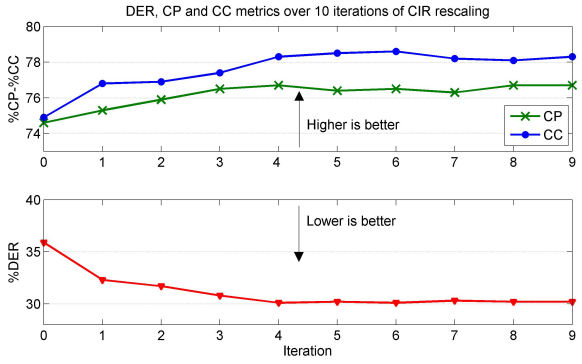

Figure 2: *System performance using iterative CIR rescaling.*

Table 2, where it can be seen that a 16% relative improvement is achieved compared to the baseline performance, with respect to the AER metric, using our CIR rescaling scheme. As before CP and CC metrics, as well as the number of unique speakers obtained, have improved as a result of CIR rescaling.

4.2. Discussion

Our proposed iterative CIR rescaling scheme can be used to achieve a form of pairwise cluster score normalisation, by taking into account the entire clustering tree and rescaling cluster scores based on the role that they play in forming this tree. This normalisation ensures that the scores belonging to clusters with high pairwise score variation, which do not display a strong relationship with any particular cluster when taking into account the clustering tree, are suppressed as a result of CIR rescaling.

5. Conclusions

We proposed a novel approach for rescaling cluster pairwise scores in hierarchical speaker clustering in order to achieve robustness in the context of speaker diarization and speaker linking. We used a state-of-the-art speaker attribution (diarization and linking) baseline system with JFA modeling, CLR scoring and complete-linkage clustering. We then proposed employing the entire hierarchical relationship between clusters as additional information to compute a novel cluster impact ratio (CIR), prior to making a clustering decision. The CIR expresses the impact of each cluster on the entire clustering process. We thus compute the CIR for each cluster and rescale the pairwise scores for that cluster using this ratio. We show that this approach can be applied to hierarchical clustering in an iterative manner and demonstrate a relative improvement of 16% in diarization performance over the SAIVT-BNEWS corpus, after only 4 iterations of CIR rescaling.

6. Acknowledgements

This research was supported by an Australian Research Council (ARC) Linkage Grant (No: LP130100110) and the Cooperative Research Centre for Smart Services.

7. References

[1] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," IBM TJ Watson Research Center, Yorktown Heights, NY, Tech. Rep., 1998.

[2] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. Springer Berlin / Heidelberg, 2008.

[3] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE ASLP*, vol. 14, no. 5, pp. 1505 –1512, 2006.

[4] D. A. V. Leeuwen, "Speaker linking in large data sets," in *Odyssey2010*, Brno, Czech Republic, June 2010, pp. 202–208.

[5] H. Bourlard, M. Ferras, N. Pappas, A. Popescu-Belis, S. Renals, F. McInnes, P. Bell, S. Ingram, and M. Guillemot, "Processing and linking audio events in large multimedia archives: The eu inevent project," in *Proceedings of SLAM 2013*, 2013.

[6] M. Ferras and H. Bourlard, "Speaker diarization and linking of large corpora," in *IEEE SLT Workshop 2012*, Dec., pp. 280–285.

[7] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *IEEE ICASSP2012*, march 2012, pp. 4185 –4188.

[8] D. Charlet, C. Fredouille, G. Damnati, and G. Senay, "Improving speaker identification in tv-shows using person name detection in overlaid text and speech," in *INTERSPEECH2013*, 2013.

[9] A. Giraudel, M. Carr, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus : a multimodal corpus for person recognition," in *Proceedings of LREC'12*, Istanbul, Turkey, may 2012.

[10] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *INTERSPEECH*, 2009, pp. 1559–1562.

[11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions ASLP*, vol. 15, no. 4, pp. 1435 –1447, may 2007.

[12] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech 2008*, 2008, pp. 853–856.

[13] C. Vaquero, A. Ortega, and E. Lleida, "Partitioning of two-speaker conversation datasets," in *Interspeech 2011*, August 28-31 2011, pp. 385–388.

[14] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Extending the task of diarization to speaker attribution," in *Interspeech2011*, Florence, Italy, August 2011. [Online]. Available: http://eprints.qut.edu.au/43351/

[15] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker linking using complete-linkage clustering," in *SST2012*, 2012.

[16] ——, "Speaker attribution of australian broadcast news data," in *SLAM2013*. Marseille, France: Sun SITE Central Europe, August 2013, pp. 72–77. [Online]. Available: http://eprints.qut.edu.au/63498/

[17] A. Jain, A. Topchy, M. Law, and J. Buhmann, "Landscape of clustering algorithms," in *Proceedings of ICPR2004*, vol. 1, 2004, pp. 260 – 263 Vol.1.

[18] P. Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[19] (2007) The NIST rich transcription website. http://www.nist.gov/speech/tests/rt/.

[20] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Odyssey2001*, June 18-22 2001, pp. 213–218.

[21] "The NIST year 2004 speaker recognition evaluation plan," NIST, Tech. Rep., 2004.